Heliyon 6 (2020) e05732

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

# On linear dimension reduction based on diagonalization of scatter matrices for bioinformatics downstream analyses

Daniel Fischer [a,*], Klaus Nordhausen [b], Hannu Oja [c]

[a] Natural Resources Institute Finland (Luke), Applied Statistical Methods, Myllytie 1, 31600 Jokionen, Finland
[b] CSTAT - Computational Statistics, Institute of Statistics & Mathematical Methods in Economics, Vienna University of Technology, Wiedner Hauptstraße 7, A-1040 Vienna, Austria
[c] Department of Mathematics and Statistics, University of Turku, 20014 Turku, Finland

ARTICLE INFO

*Keywords:*
Computer science
Mathematics
Statistics
Bioinformatics
Microbial genomics
Genomics
Transcriptomics
Dimension reduction
SIR
ICA

ABSTRACT

Dimension reduction is often a preliminary step in the analysis of data sets with a large number of variables. Most classical, both supervised and unsupervised, dimension reduction methods such as principal component analysis (PCA), independent component analysis (ICA) or sliced inverse regression (SIR) can be formulated using one, two or several different scatter matrix functionals. Scatter matrices can be seen as different measures of multivariate dispersion and might highlight different features of the data and when compared might reveal interesting structures. Such analysis then searches for a projection onto an interesting (signal) part of the data, and it is also important to know the correct dimension of the signal subspace. These approaches usually make either no model assumptions or work in wide classes of semiparametric models. Theoretical results in the literature are however limited to the case where the sample size exceeds the number of variables which is hardly ever true for data sets encountered in bioinformatics. In this paper, we briefly review the relevant literature and explore if the dimension reduction tools can be used to find relevant and interesting subspaces for small-$n$-large-$p$ data sets. We illustrate the methods with a microarray dataset of prostate cancer patients and healthy controls.

## 1. Introduction

In contemporary data analysis linear dimension reduction is often the first step in reducing the number of variables. For a numeric $p$-variate random vector $\boldsymbol{x}$ this means that a $q \times p$ transformation matrix $\boldsymbol{W}$ with $q \ll p$ is searched such that all relevant information for the analysis at hand is contained in $\boldsymbol{Wx}$. The question is then naturally how "information" content is measured but in general two types of linear dimension reduction can be distinguished:

**Unsupervised Dimension Reduction:**
$\boldsymbol{x}|\boldsymbol{Wx}$ is considered as uninformative noise.
**Supervised Dimension Reduction:**
$y \perp\!\!\!\perp \boldsymbol{x}|\boldsymbol{Wx}$ for some response variable $y$ of interest.

There is an abundance of supervised and unsupervised methods, and depending on the data and problem, they might give quite different results as they adopt different concepts of information. One prevalent supervised method is, e.g. the linear discriminant analysis (LDA) or more recently developed non-linear dimension reduction methods like t-SNE (van der Maaten and Hinton, 2008), Isomap (Tenenbaum et al., 2000) or UMAP (McInnes et al., 2018).

However, in the context of linear unsupervised dimension reduction, for example, principal component analysis (PCA) understands information as a large variation whereas independent component analysis (ICA) usually measures information as a degree of non-gaussianity. The dimension $q$ of the information or signal subspace is preselected visually or by using various information criteria or testing strategies. For a more detailed comparison of the two concepts with practical examples, please see Nordhausen and Oja (2018b).

Surprisingly many unsupervised and supervised linear dimension reduction methods can be expressed as a joint diagonalization problem of two scatter matrices which yields a nice unifying theory in wide semiparametric models. A problem, however, is that the assumption $n > p$ for the sample size $n$ is needed and the technique cannot be directly applied to bioinformatics data with almost always $n \ll p$.

In this paper, we will first provide in Section 2 a general but brief overview of linear dimension reduction methods based on the use of

* Corresponding author.
*E-mail address:* daniel.fischer@luke.fi (D. Fischer).

https://doi.org/10.1016/j.heliyon.2020.e05732
Received 18 January 2020; Received in revised form 1 June 2020; Accepted 11 December 2020

two scatter matrices. Section 3 introduces a genetic data set which is used to illustrate the ideas. Section 4 then explores if there are any ways how to apply these methods in the case $n \ll p$ with the example data discussed in Section 3.

A GitHub repository that contains the R-code and data to reproduce all results and figures is also available (https://github.com/fischuu/LinearDimensionReductionInBioinformatics).

## 2. Linear dimension reduction using two of scatter matrices

Let $\boldsymbol{x}$ denote a $p$-variate random vector having cdf $F_{\boldsymbol{x}}$ and the joint cdf of random variable $y$ and $\boldsymbol{x}$ is denoted by $F_{\boldsymbol{x},y}$. An (unsupervised) scatter functional is then any $p \times p$ matrix valued functional $\boldsymbol{S}$ which is affine equivariant in the sense that

$$\boldsymbol{S}(F_{\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}}) = \boldsymbol{A}\boldsymbol{S}(F_{\boldsymbol{x}})\boldsymbol{A}^{\top}$$

for all nonsingular $p \times p$ matrices $\boldsymbol{A}$ and all $p$-vectors $\boldsymbol{b}$. Similarly, a supervised scatter functional is defined as any $p \times p$ matrix valued functional $\boldsymbol{S}$ which is affine equivariant in the sense that

$$\boldsymbol{S}(F_{\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b},y}) = \boldsymbol{A}\boldsymbol{S}(F_{\boldsymbol{x},y})\boldsymbol{A}^{\top},$$

with $\boldsymbol{A}$ and $\boldsymbol{b}$ as above. These functionals are in the following also denoted by $\boldsymbol{S}(\boldsymbol{x})$ and $\boldsymbol{S}(\boldsymbol{x}; y)$. For a random sample $\boldsymbol{X}$ (or $(\boldsymbol{X}, y)$), the estimates of the population values $\boldsymbol{S}(\boldsymbol{x})$ and $\boldsymbol{S}(\boldsymbol{x}; y)$ are obtained as the value of the functional at the corresponding empirical distributions.

There is a wide literature on scatter functionals. The covariance matrix $\mathrm{cov}(\boldsymbol{x})$ serves as the first example but the scatter matrix may also be based on the fourth moments:

$$\mathrm{cov}_4(\boldsymbol{x}) = E\left(r^2(\boldsymbol{x} - \boldsymbol{E}(\boldsymbol{x}))(\boldsymbol{x} - \boldsymbol{E}(\boldsymbol{x}))^{\top}\right),$$

with $r^2 = (\boldsymbol{x} - E(\boldsymbol{x}))^{\top}\mathrm{cov}(\boldsymbol{x})^{-1}(\boldsymbol{x} - E(\boldsymbol{x}))$. The M-functionals of scatter are implicitly defined by

$$\boldsymbol{S}(\boldsymbol{x}) = E\left(w(r)(\boldsymbol{x} - \boldsymbol{T}(\boldsymbol{x}))(\boldsymbol{x} - \boldsymbol{T}(\boldsymbol{x}))^{\top}\right),$$

where $\boldsymbol{T}(\boldsymbol{x})$ is a companion location functional and $w(\cdot)$ a weight function for $r = ((\boldsymbol{x} - \boldsymbol{T}(\boldsymbol{x}))^{\top}\boldsymbol{S}(\boldsymbol{x})^{-1}(\boldsymbol{x} - \boldsymbol{T}(\boldsymbol{x})))^{1/2}$. Popular weight functions are for example Huber weights

$$w_H(r) = \begin{cases} 1/\sigma^2 & r \le c \\ c/(r^2\sigma^2) & r > c \end{cases},$$

where $c$ is an user specified tuning constant and $\sigma^2$ a scaling factor. The weight function for Tyler's scatter matrix is $w_T(r) = p/r^2$. For a general discussion on these and other (unsupervised) scatter matrices, see for example Tyler (1987); Rousseeuw and Hubert (2013); Dümbgen et al. (2015). A supervised scatter functional used in the sliced inverse regression (SIR) is

$$\boldsymbol{S}_{SIR}(\boldsymbol{x}; y) = E((\boldsymbol{x} - E(\boldsymbol{x}|y))(\boldsymbol{x} - E(\boldsymbol{x}|y))^{\top}). \quad (1)$$

For more examples of supervised functionals, see, e.g. Liski et al. (2014).

Initially scatter functionals were developed to provide robust and efficient competitors to the covariance matrix under the multivariate normality or ellipticity assumptions. In the elliptic models, the scatter functionals can be shown to be proportional to the covariance matrix (see, e.g. Nordhausen et al., 2011). A property of interest is also the so-called independence property (Nordhausen and Tyler, 2015) which states that if $\boldsymbol{x}$ has independent components, then $\boldsymbol{S}(\boldsymbol{x})$ is a diagonal matrix. The covariance matrix, the scatter matrix based on the fourth moments as well as symmetrized versions of Huber's and Tyler's M-estimates all have this property, see for example Dümbgen (1998); Sirkiä et al. (2007).

Surprisingly many linear dimension reduction methods can be seen as a joint diagonalization of two scatter matrices. Let $\boldsymbol{W} = \boldsymbol{W}(\boldsymbol{x})$ be a $p \times p$ transformation matrix (functional) which diagonalizes two scatter functionals $\boldsymbol{S}_1 = \boldsymbol{S}_1(\boldsymbol{x})$ and $\boldsymbol{S}_2 = \boldsymbol{S}_2(\boldsymbol{x})$ so that

$$\boldsymbol{W}\boldsymbol{S}_1\boldsymbol{W}^{\top} = \boldsymbol{I}_p \quad \text{and} \quad \boldsymbol{W}\boldsymbol{S}_2\boldsymbol{W}^{\top} = \boldsymbol{D},$$

where $\boldsymbol{D}$ is a $p \times p$ diagonal matrix with diagonal elements $d_1 \ge \ldots \ge d_p$. This method is called invariant coordinate selection (ICS) (Tyler et al., 2009) as, for any nonsingular $p \times p$ matrix $\boldsymbol{A}$, $\boldsymbol{W}(\boldsymbol{x})(\boldsymbol{x}) = \boldsymbol{W}(\boldsymbol{A}\boldsymbol{x})\boldsymbol{A}\boldsymbol{x}$ up to the signs of the components. The transformed $\boldsymbol{W}(\boldsymbol{x})\boldsymbol{x}$ in an invariant coordinate system may reveal intrinsics and hidden structures in the data. For unsupervised $\boldsymbol{S}_1$ and supervised $\boldsymbol{S}_2$, the transformation is known as supervised invariant coordinate selection (SICS) (Liski et al., 2014). ICS and SICS are very general methods with many several possible applications - but now have a look only at three well-known special cases.

### 2.1. Principal component analysis

PCA is one of the most popular multivariate methods and fits into this framework with $\boldsymbol{S}_1 := \boldsymbol{I}_p$ and $\boldsymbol{S}_2 := \mathrm{cov}$. The principal components in $\boldsymbol{W}(\boldsymbol{x})\boldsymbol{x}$ are then ordered according to their variances, that is, the diagonal elements of $\boldsymbol{D}$. The number of components $q$ to choose is often based on the cumulative proportion of variation or, visually, finding the elbow in the scree-plot indicating that the variances of the remaining components are approximately equal.

The idea of the scree-plot checking is formalized by assuming that, for some $q$, the principal values satisfy $d_1 \ge \ldots \ge d_q > d_{q+1} = \ldots = d_p > 0$. The null hypothesis $H_0 : q = k$ can then be tested by assuming ellipticity and using the variance of the $p - k$ smallest estimated eigenvalues, say $T_k$, as a test statistic. The limiting distribution of $nT_q$, properly scaled and when $n \to \infty$, is then a chi squared distribution with $(p - q - 1)(p - q + 2)/2$ degrees of freedom. For asymptotic and bootstrap tests and the estimates of $q$ based on these tests, see Schott (2006); Nordhausen et al. (2017a).

### 2.2. Independent component analysis

The independent component model is a semiparametric model with the assumption that

$$\boldsymbol{x} = \boldsymbol{A}\boldsymbol{z} + \boldsymbol{b},$$

where $\boldsymbol{A}$ is a full-rank $p \times p$ mixture matrix, $\boldsymbol{b}$ specifies the location center of $\boldsymbol{x}$, and $\boldsymbol{z}$ is a standardized random $p$-vector of independent components so that $E(\boldsymbol{z}) = \boldsymbol{0}$ and $\mathrm{cov}(\boldsymbol{z}) = \boldsymbol{I}_p$. The goal in ICA is to estimate an unmixing matrix $\boldsymbol{W}$ to transform the data to the independent components (Nordhausen and Oja, 2018a).

The matrix $\boldsymbol{W}$ also solves the ICA problem if both scatter matrices $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$ have the independence property. The eigenvalues in $\boldsymbol{D}$ provide componentwise kurtosis measures (Nordhausen et al., 2011). The most popular choice providing the FOBI solution is $\boldsymbol{S}_1 := \mathrm{cov}$ and $\boldsymbol{S}_2 := \mathrm{cov}_4$ with the componentwise Pearson's kurtosis measures (see Cardoso (1989) and also Nordhausen and Virta (2019)). The $q$ non-gaussian independent components can be estimated in this model if their kurtosis values are distinct. In signal processing, it is generally thought that these $q$ non-gaussian components present the signal part and the $p - q$ gaussian components the noise part of the data.

For the FOBI approach, the eigenvalues for the gaussian components are $p + 2$ and, to find the non-gaussian components, the strategy is to choose the components whose values differ most from $p + 2$. A natural test statistic $T_k$ for $H_0 : q = k$ is then the sum of $p - k$ smallest diagonal entries of $(\widehat{\boldsymbol{D}} - (p + 2)\boldsymbol{I}_p)^2$ and, if the eight moments are finite, then the limiting null distribution of $nT_q$ as $n \to \infty$ is the distribution of $2\sigma_1\chi^2_{\frac{1}{2}(p-q-1)(p-q+2)} + \left(2\sigma_1 + 4(p - q)\right)\chi^2_1$ with independent chi squared variables. The constant $\sigma_1 = Var\left(\|\boldsymbol{z}\|^2\right) + 8$ can be consistently estimated from the data. The bootstrap test versions are also

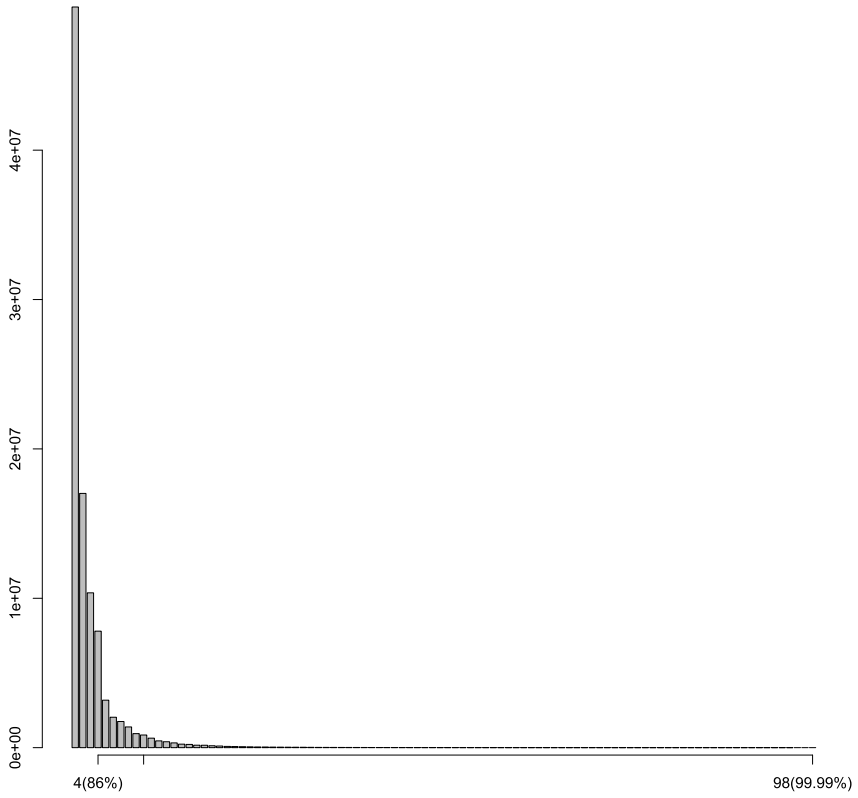


**Fig. 1.** Scree-plot for the squared singular values.

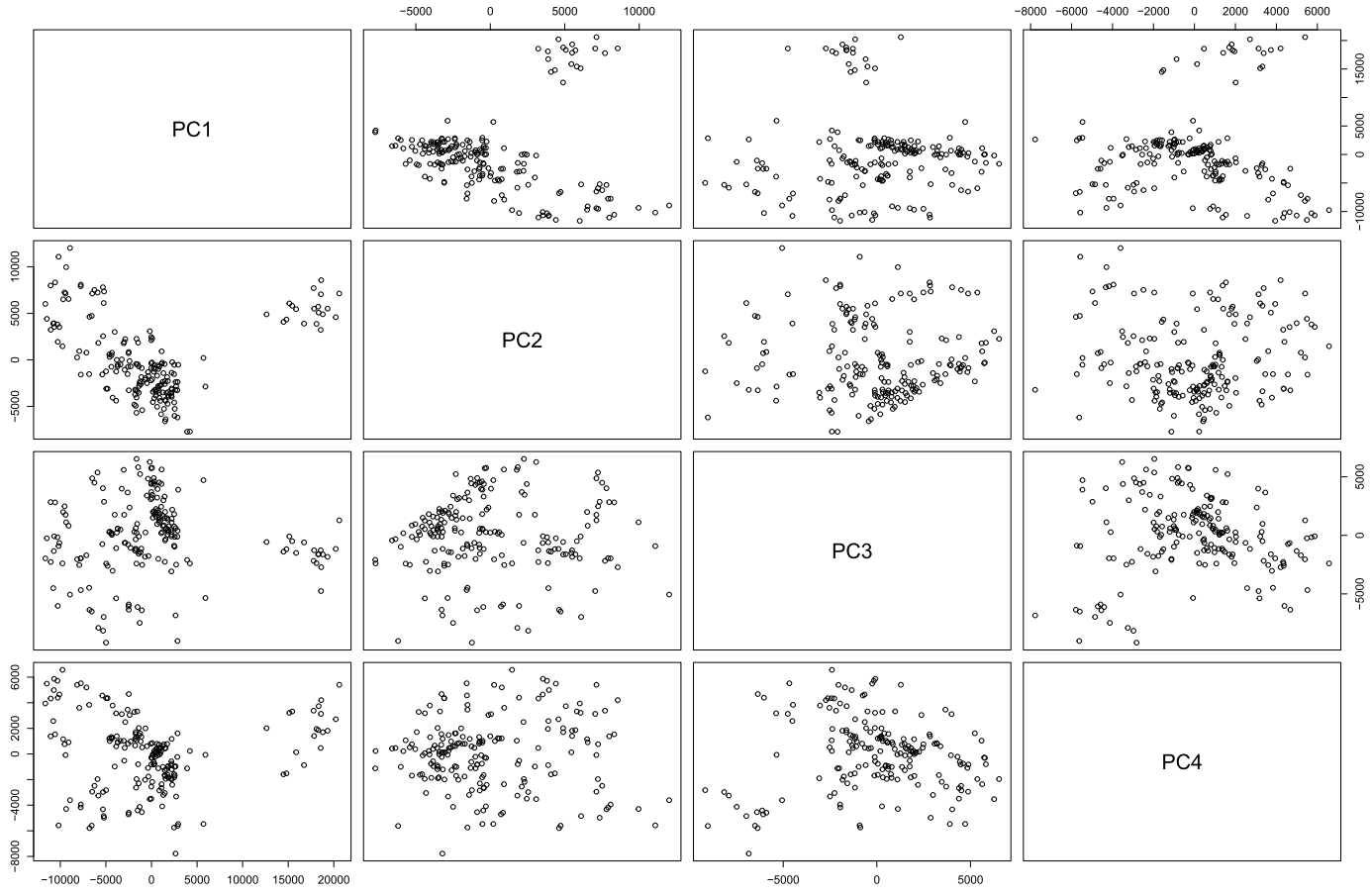


**Fig. 2.** Pairwise scatter plots for the first four principal components.

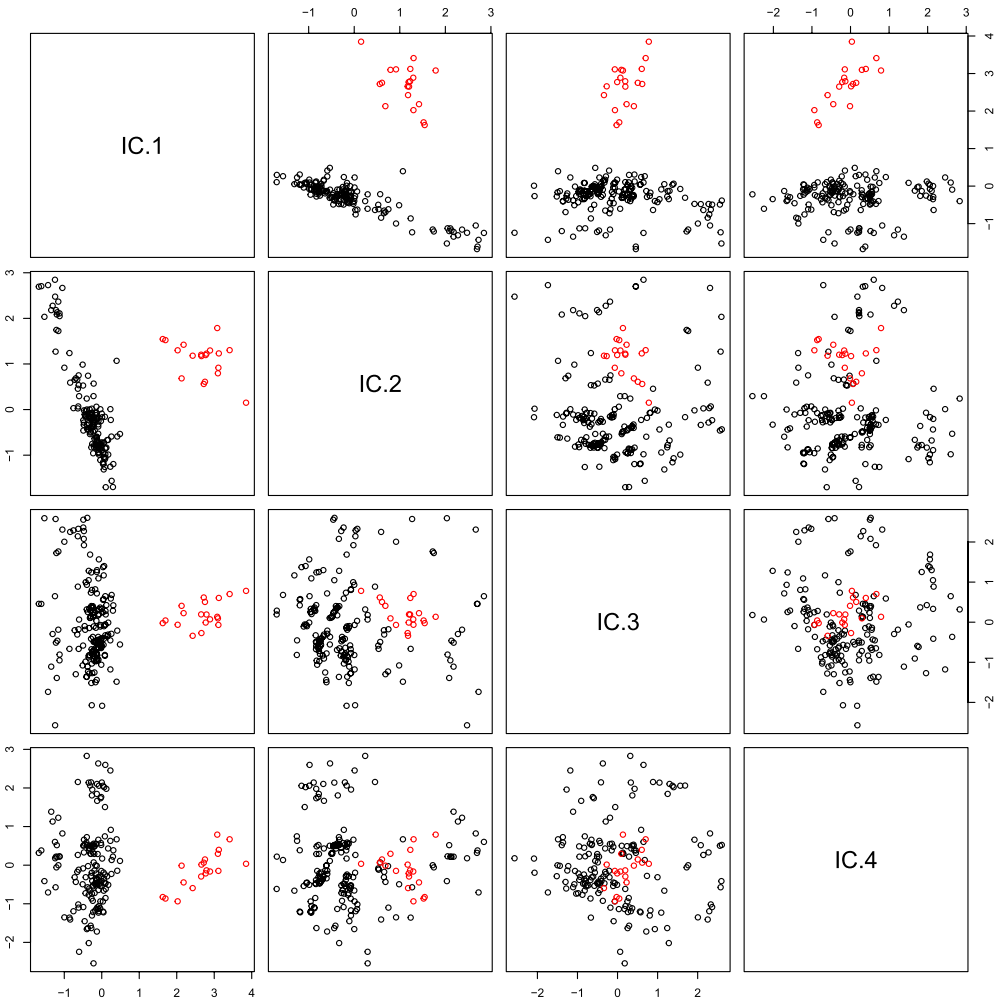


**Fig. 3.** Independent components based on FOBI.

easily available. Consistent estimates of $q$ can be found using various sequential testing strategies. Further, the ICA assumptions can be relaxed by allowing that the non-gaussian components are dependent; this is known as non-gaussian component analysis (NGCA). For these and further results, see Nordhausen et al. (2017a) and Nordhausen et al. (2017b).

### 2.3. Sliced inverse regression

SIR is a supervised dimension reduction method originally suggested by Li (1991). It uses $\boldsymbol{S}_1 := \text{cov}$ and $\boldsymbol{S}_2 = \boldsymbol{S}_{SIR}$ as defined in equation (1). In practice, one discretizes $y$ and uses $\boldsymbol{S}_{SIR}(\boldsymbol{x}, y^d)$ where $y^d = h \Leftrightarrow y \in S_h$, $h = 1, ..., H$, with disjoint intervals (slices) $S_1, ..., S_H$ which satisfy $S_1 + ... + S_H = \mathbb{R}$. Note that the rank of the sample version of $\boldsymbol{S}_{SIR}(\boldsymbol{x}, y^d)$ is at most $H - 1$. We then assume again the location-scatter model

$$\boldsymbol{x} = \boldsymbol{A}\boldsymbol{z} + \boldsymbol{b},$$

where now the standardized $\boldsymbol{z} = (\boldsymbol{z}_1^\top, \boldsymbol{z}_2^\top)^\top$ satisfies $(y, \boldsymbol{z}_1^\top)^\top \perp\!\!\!\perp \boldsymbol{z}_2$. In the partitioning, $\boldsymbol{z}_1$ is chosen to have the smallest possible dimension $q$. The subvectors $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ present the signal and noise parts of $\boldsymbol{z}$, respectively. Under this model and using $\text{cov}(\boldsymbol{x})$ and $\boldsymbol{S}_{SIR}(\boldsymbol{x}, y^d)$, the diagonal entries of $\boldsymbol{D}$ are

$$d_1 \geq \ldots \geq d_q \geq d_{q+1} = \ldots = d_p = 0.$$

It is further required that $d_q > d_{q+1}$ so that SIR is assumed to find the full signal.

To test the hypothesis $H_0 : q = k$ Nordhausen et al. (2017a) used the test statistic $T_k$ which is simply the sum of the $p - k$ smallest diagonal entries of $\widehat{\boldsymbol{D}}$. Then, for the true value $q$, $nT_q$ has a limiting $\chi^2_{(p-q)(H-q-1)}$ distribution as $n$ goes to the infinity. As in the PCA and FOBI cases, different sequential testing can be again used to find a consistent estimate of $q$. Also, for small $n$, a bootstrap testing strategy may be used. For these and further results, see again Nordhausen et al. (2017a). Bura and Cook (2001) derived the limiting distribution of the same test statistic under weaker conditions.

### 2.4. Comparison to other dimension reduction methods

The here considered linear dimension reduction methods rely on linear transformations of the original data into a lower dimensional subspace. In contrast, non-linear dimension reduction methods like t-SNE, Isomap and UMAP apply more drastical transformations to the data. These methods do not assume that the original data lies on a linear subspace.

Besides linear/non-linear methods for dimension reduction, even feature selection methods several classification methods like, e.g. the Random Forest can be considered as a dimension reduction method. Also, even a simple linear regression can be used for dimension reduction.

### 2.5. The case of $n \ll p$

The use of the above methodology (PCA, FOBI, SIR) involves problems in the context of genetic data with $n \ll p$. Tyler (2010) showed that

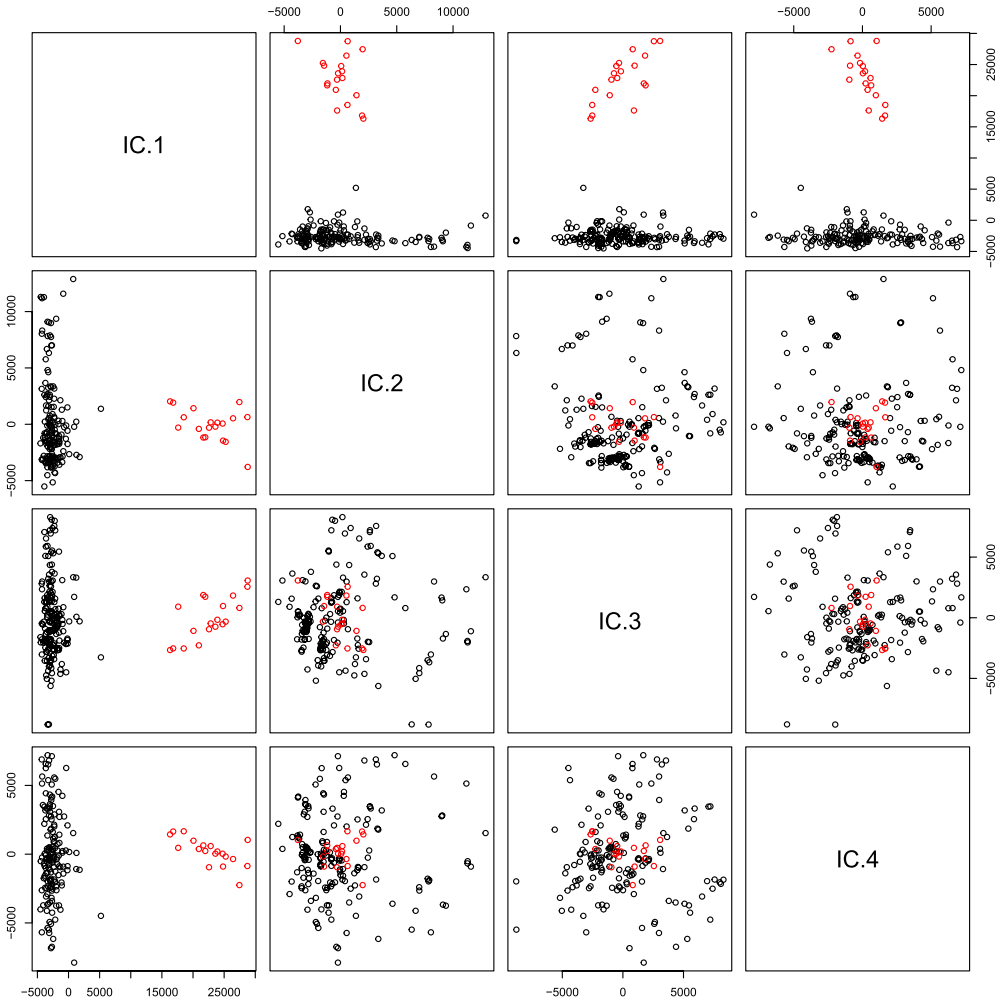


**Fig. 4.** Independent components based on robust ICA.

in this scenario, unfortunately, all scatter functionals with the affine equivariance property are proportional to the covariance matrix prohibiting the subspace estimation. So to apply these methods, one should first reduce the dimension to some $k < n$ by using SVD for example and then continue working in the resulting subspace.

The high dimensionality problem is approached in bioinformatics often by performing PCA via the singular value decomposition (SVD). This is the quasi-standard first step in bioinformatics data analysis, see the tools such as Genomatix, Genious or CLCbio that "solve" the problem of choosing a working dimension $k$ quite conveniently and simply by setting the default to $k = 2$. Let $\boldsymbol{X}$ be our centered $p \times n$ data matrix, then the singular value decomposition (SVD) for $\boldsymbol{X}$ with rank $r \leq \min\{p, n\}$ is defined as

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\top}$$

where $\boldsymbol{U} = (\boldsymbol{u}_1, ..., \boldsymbol{u}_r)$ is a $p \times r$ matrix and $\boldsymbol{V} = (\boldsymbol{v}_1, ..., \boldsymbol{v}_r)$ is an $n \times r$ matrix both with orthonormal columns, and $\boldsymbol{D}$ is an $r \times r$ diagonal matrix with positive diagonal elements in a decreasing order. The number of variables via PCA is then reduced to $k$ with a transformation $\boldsymbol{X}^* = (\boldsymbol{U}^*)'\boldsymbol{X}$ where $\boldsymbol{U}^* = (u_1, ..., u_k)$.

The problem then naturally is how to choose $k$ without losing any (or too much) information.

## 3. A microRNA expression data set

Analysis of genetic data sets is one of the driving forces behind developing the tools for the $n \ll p$ problems. For example, the current sample sizes in typical transcriptomic experiments range from just a few to a couple of hundreds of individuals due to the high costs of sequencing. Also, the storage of massive data is often almost prohibitive. It is therefore not likely that in the foreseeable future, the case $n > p$ would be realistic for genetic data.

The data set used in this work consists of human microRNA (miRNA) expressions from Agilent microarrays where 2125 different probes of 813 different miRNAs were used for subjects coming from three different groups: 76 healthy individuals, 78 patients with a mild form of prostate cancer (PrCa) and 35 with an aggressive type of PrCa. In total, 26 microarrays have been used, and the hybridization took place at four different time points. The data set was originally analyzed in Fischer et al. (2014, 2015), and all relevant data are available from EMBL-EBI ArrayExpress (accession number E-MTAB-3397). Note that here $n = 189$ may be seen untypically large compared to the relatively "small" dimension $p = 2125$. The goal in this data is to identify miRNAs which are either responsible for the development of PrCa (predisposition) or which could serve as biomarkers for the detection of PrCa (diagnosis) and, further, to distinguish the two different types of PrCa.

Also, please note that the here described findings are not only valid for microarray data but also for any other kind of more recent data like e.g. whole-transcriptome sequencing (WTS). In WTS the $n << p$ problem is even more prominent as, compared to the microarray data, usually all possible genes respective transcripts are quantified so that $n$ is in the level of magnitude $> 20.000$ and $p$ is often also smaller. Hence, also in WTS dimension reduction methods are eminent.

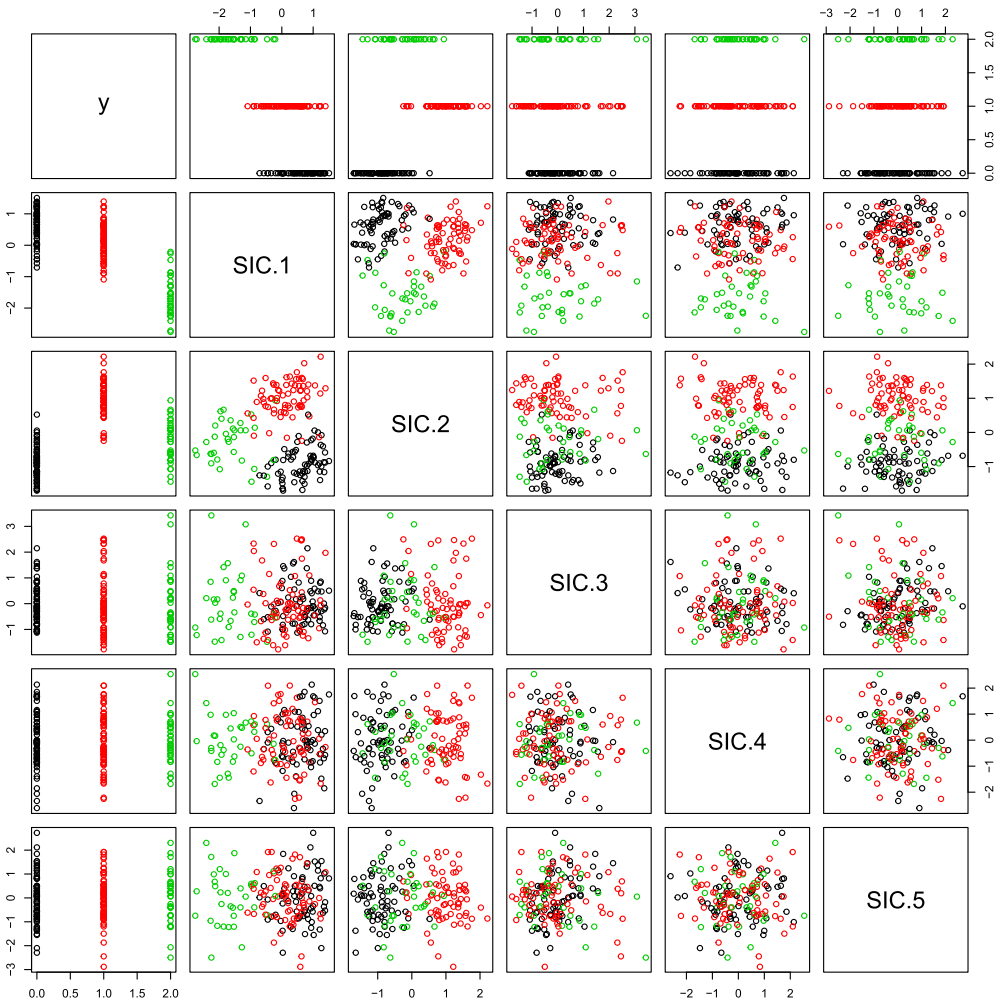


**Fig. 5.** Pairwise scatter plots for SIR components obtained from the 98-variate data and with the colors corresponding to the three health groups.

**Table 1**. The first 10 squared singular values from SVD.

| | Squared singular value | Cumulative explained variation (%) |
|---|---|---|
| 1 | 49584782.29 | 50.3% |
| 2 | 17020792.70 | 67.6% |
| 3 | 10365087.69 | 78.1% |
| 4 | 7799839.61 | 86.0% |
| 5 | 3178859.60 | 89.2% |
| 6 | 2036165.39 | 91.3% |
| 7 | 1750698.70 | 93.1% |
| 8 | 1386901.81 | 94.5% |
| 9 | 937702.68 | 95.5% |
| 10 | 846675.15 | 96.3% |

## 4. Application

In the analysis of the microRNA data, the problem was to identify those miRNAs which allow us to separate healthy subjects from PrCa cases and, if possible, even distinguish between the two different types of cancer. The following analysis was done entirely with R (R Core Team, 2017) using the packages ICS (Nordhausen et al., 2008) and ICtest (Nordhausen et al., 2017c).

As $n \ll p$, the SVD was performed with the results reported in Table 1 (the first 10 eigenvalues of cov) and the corresponding scree-plot in Fig. 1. The scree-plot suggests that four components might be reasonable as they already explain more than 80% of variation of the data (which has also sometimes been used as a rule). Note that as many as 98 components are needed to explain 99.99% of the variation.

**Table 2**. Ordered kurtosis values from FOBI and robust ICA using the first four principal components.

| | FOBI | robust ICA |
|---|---|---|
| 1 | 9.7428 | 3.6838 |
| 2 | 7.2991 | 0.8069 |
| 3 | 5.3417 | 0.6101 |
| 4 | 6.2365 | 0.5514 |

*ICA in the four-variate subspace* We first start with four principal components ($k = 4$) plotted in Fig. 2. The pairwise scatter plots for these four first principal components reveal one group that is separate from the bulk of the data but not three groups with different cancer types as desired. This is because, in searching for subgroups of the data, the kurtosis measures are more natural than the variance as an information criterion. We therefore next try ICS for this four-variate data with two choices of the pairs of scatter matrices $(\boldsymbol{S}_1, \boldsymbol{S}_2)$, namely, (i) FOBI based on cov and $\text{cov}_4$, and (ii) robust ICA based on symmetrized versions Tyler's and Huber's scatter matrices. Table 2 describes the estimated kurtosis measures and the diagonal entries of $\widehat{\boldsymbol{D}}$.

The test results reported in Table 3 suggest that the number of non-gaussian components is $q = 1$. The four ICS components from FOBI and robust ICA are plotted in Figs. 3 and 4, respectively. In both cases, the first component seems to separate the two groups, which unfortunately are not at all connected to the three health status groups. After some detective work, it was found out that the group of 24 subjects, highlighted with red color in the plots, were outliers from three microarrays created

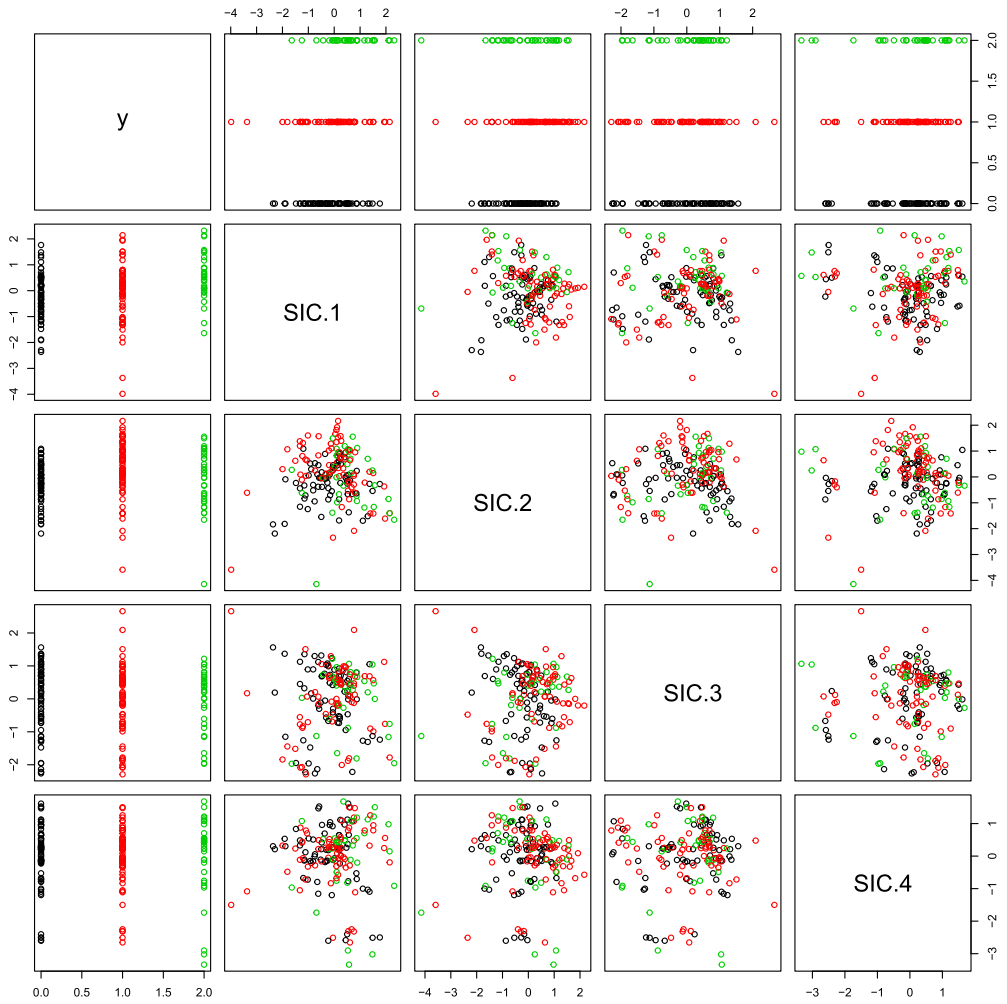


**Fig. 6.** Pairwise scatter plots for SIR components obtained from the four-variate data and with the colors corresponding to the three health groups.

**Table 3**. The $p$-values for testing the hypotheses of $H_0 : q = k$, $k = 0, 1, 2$ where $q$ the number of non-gaussian components. The tests are asymptotic and bootstrap tests (with 500 bootstrap samples) using the FOBI approach for the four-variate data.

| $H_0$ | Asymp | Boot |
|---|---|---|
| $q = 0$ | <0.0001 | 0.0020 |
| $q = 1$ | 0.1050 | 0.0818 |
| $q = 2$ | 0.3963 | 0.4711 |

by a single person at different time points. Also, the second components seem just to find another small group of outliers in the data. Hence, based on these results, it can be concluded that ICA cannot find the groups of interest if only the first four principal components are used in the analysis. We repeated ICA with 98 principal components as well, but even in this case could not find structures in the data to identify the health status of the subjects.

*SIR in the 98-variate subspace* We next perform SIR and use the data with the 98 principal components. As we hope to separate the three health groups, it is natural to let $y$ indicate the group memberships. Note that, as $H = 3$, the dimension of the interesting subspace is $q = 0, 1$ or 2. The $p$-values for the asymptotic and bootstrap tests for $q = 0$ and $q = 1$ are reported in Table 4. The tests reject both $q = 0$ and $q = 1$ and therefore suggest that $q = 2$. For visualization purposes we plot, in Fig. 5, the first five components and highlight the three response classes with different colors. We also applied SIR to the four-variate data, and the SICS components are plotted in Fig. 6. It is then clear that the information to separate the three health groups is lost if only the first four principal components are used.

**Table 4**. The $p$-values for testing $H_0 : q = 0$ and $H_0 : q = 1$. The tests are asymptotic and bootstrap tests (with 500 bootstrap samples) using the SIR approach for the 98-variate data.

| $H_0$ | Asymp | Boot |
|---|---|---|
| $q = 0$ | ≤0.0001 | 0.0448 |
| $q = 1$ | 0.0001 | 0.0050 |

## 5. Conclusions

In the paper, we discussed the use of two scatter matrices for the unsupervised and supervised linear dimension reduction under broad model assumptions. The signal and noise spaces are then easily separated using the scatter matrices, and the eigenvalues of one scatter matrix with respect to another one listed in $\boldsymbol{D}$. The eigenvalues can also be used to decide what is the dimension of the signal space. The theory is however developed only for large-$n$-small-$p$ cases which rules out genetic applications.

Following the general practice in bioinformatics, we tried to circumvent this problem by reducing the dimension using SVD (PCA) and hoping that a few first principal components capture all the relevant information. In our case study, we explored this strategy and used ICA and SIR for four- and 98-variate data sets obtained in this way. ICA and SIR have been used to analyze bioinformatics data earlier also in

Liebermeister (2002); Kong et al. (2008); Zhong et al. (2005); Fischer et al. (2017). Here we, however, wanted to explore what is the impact of the SVD step on the results, and the conclusion is that the number of principal components to retain is crucial. It seems advisable to keep the number of components as high as the further analysis tools allow without suffering from the course of dimensionality.

Based on this study, it would be worthwhile to develop models and techniques which allow the SVD preprocessing step and can then formalize the rules for the number of PCs to retain. The SVD step could, for example, be incorporated into the bootstrapping procedure to accommodate the variation coming from that step. Note however that for example El Karoui and Purdom (2018, 2019) argue that bootstrapping is in general difficult in the $n \ll p$ setting.

## Declarations

### *Author contribution statement*

D. Fischer: Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper. K. Nordhausen: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper. H. Oja: Conceived and designed the experiments; Contributed reagents, materials, analysis tools or data; Wrote the paper.

### *Funding statement*

The work of KN was supported by the Austrian Science Fund (FWF) Grant number P31881-N32.

### *Data availability statement*

Data associated with this study has been deposited at https://github.com/fischuu/LinearDimensionReductionInBioinformatics.

### *Declaration of interests statement*

The authors declare no conflict of interest.

### *Additional information*

No additional information is available for this paper.

## Acknowledgements

The authors wish to acknowledge CSC – IT Center for Science, Finland, for computational resources. We thank two anonymous reviewers for helpful comments on an earlier draft of the manuscript.

## References

Bura, E., Cook, R.D., 2001. Extending sliced inverse regression: the weighted chi-squared test. J. Am. Stat. Assoc. 96, 996–1003.

Cardoso, J.F., 1989. Source separation using higher order moments. In: Proceedings of IEEE International Conference on Accoustics, Speech and Signal Processing. Glasgow, UK, pp. 2109–2112.

Dümbgen, L., 1998. On Tyler's $M$-functional of scatter in high dimension. Ann. Inst. Stat. Math. 50, 471–491.

Dümbgen, L., Pauly, M., Schweizer, T., 2015. $M$-functionals of multivariate scatter. Stat. Surv. 9, 32–105.

El Karoui, N., Purdom, E., 2018. Can we trust the bootstrap in high-dimensions? The case of linear models. J. Mach. Learn. Res. 19, 1–66.

El Karoui, N., Purdom, E., 2019. The bootstrap, covariance matrices and pca in moderate and high-dimensions. In: Proceedings of Machine Learning Research: the 22nd International Conference on Artificial Intelligence and Statistics, vol. 89, pp. 2115–2124.

Fischer, D., Honkatukia, M., Tuiskula-Haavisto, M., Nordhausen, K., Cavero, D., Preisinger, R., Vilkki, J., 2017. Subgroup detection in genotype data using invariant coordinate selection. BMC Bioinform. 18, 173–181.

Fischer, D., Oja, H., Schleutker, J., Sen, P.K., Wahlfors, T., 2014. Generalized Mann-Whitney type tests for microarray experiments. Scand. J. Stat. 41, 672–692.

Fischer, D., Wahlfors, T., Mattila, H., Oja, H., Tammela, T., Schleutker, J., 2015. Mirna profiles in lymphoblastoid cell lines of Finnish prostate cancer families. PLoS ONE 10, e0127427.

Kong, W., Vanderburg, C., Gunshin, H., Rogers, J., Huang, X., 2008. A review of independent component analysis application to microarray gene expression data. BioTechniques 45, 501–520.

Li, K.-C., 1991. Sliced inverse regression for dimension reduction. J. Am. Stat. Assoc. 86, 316–327.

Liebermeister, W., 2002. Linear modes of gene expression determined by independent component analysis. Bioinformatics 18, 51–60.

Liski, E., Nordhausen, K., Oja, H., 2014. Supervised invariant coordinate selection. Statistics 48, 711–731.

van der Maaten, L., Hinton, G., 2008. Visualizing data using t-sne. J. Mach. Learn. Res. 9, 2579–2605.

McInnes, L., Saul, J.H., Grossberger, L., 2018. Umap: uniform manifold approximation and projection. J. Open Sour. Softw. 3, 2861.

Nordhausen, K., Oja, H., 2018a. Independent Component Analysis: A Statistical Perspective. Wiley Interdisciplinary Reviews: Computational Statistics, vol. 10, p. e1440.

Nordhausen, K., Oja, H., 2018b. Robust nonparametric inference. Annu. Rev. Stat. Appl. 5, 473–500.

Nordhausen, K., Oja, H., Ollila, E., 2011. Multivariate models and the first four moments. In: Hunter, D.R., Richards, D.S.R., Rosenberger, J.L. (Eds.), Nonparametric Statistics and Mixture Models. World Scientific, Singapore, pp. 267–287.

Nordhausen, K., Oja, H., Tyler, D.E., 2008. Tools for exploring multivariate data: the package ICS. J. Stat. Softw. 28, 1–31.

Nordhausen, K., Oja, H., Tyler, D.E., 2017a. Asymptotic and bootstrap tests for subspace dimension. ArXiv e-prints.

Nordhausen, K., Oja, H., Tyler, D.E., Virta, J., 2017b. Asymptotic and bootstrap tests for the dimension of the non-Gaussian subspace. IEEE Signal Process. Lett. 24, 887–891.

Nordhausen, K., Oja, H., Tyler, D.E., Virta, J., 2017c. ICtest: estimating and testing the number of interesting components in linear dimension reduction. https://CRAN.R-project.org/package=ICtest. R package version 0.3.

Nordhausen, K., Tyler, D.E., 2015. A cautionary note on robust covariance plug-in methods. Biometrika 102, 573–588.

Nordhausen, K., Virta, J., 2019. An overview of properties and extensions of FOBI. Knowl.-Based Syst. 173, 113–116.

R Core Team, 2017. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Rousseeuw, P., Hubert, M., 2013. High-breakdown estimators of multivariate location and scatter. In: Becker, C., Fried, R., Kuhnt, S. (Eds.), Robustness and Complex Data Structures: Festschrift in Honour of Ursula Gather. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 49–66.

Schott, J.R., 2006. A high-dimensional test for the equality of the smallest eigenvalues of a covariance matrix. J. Multivar. Anal. 97, 827–843.

Sirkiä, S., Taskinen, S., Oja, H., 2007. Symmetrised $M$-estimators of scatter. J. Multivar. Anal. 98, 1611–1629.

Tenenbaum, J.B., de Silva, V., Langford, J.C., 2000. A global geometric framework for nonlinear dimensionality reduction. Science 290, 2319–2323.

Tyler, D.E., 1987. A distribution-free $M$-estimator of multivariate scatter. Ann. Stat. 15, 234–251.

Tyler, D.E., 2010. A note on multivariate location and scatter statistics for sparse data sets. Stat. Probab. Lett. 80, 1409–1413.

Tyler, D.E., Critchley, F., Dümbgen, L., Oja, H., 2009. Invariant coordinate selection. J. R. Stat. Soc. B 71, 549–592.

Zhong, W., Zeng, P., Ma, P., Liu, J.S., Zhu, Y., 2005. Rsir: regularized sliced inverse regression for motif discovery. Bioinformatics 21, 4169–4175.